

FOR

METHOD FOR ASSIGNING A STREAMING MEDIA SESSION TO A SERVER IN FIXED AND MOBILE STREAMING MEDIA SYSTEMS

Inventors:
JOHN G. APOSTOLOPOULOS
SUJOY BASU
GENE CHEUNG
RAJ KUMAR
SUMIT ROY
BO SHEN
WAI-TIAN TAN
SUSIE J. WEE
TINA WONG

Prepared by:
WAGNER, MURABITO & HAO LLP
Two North Market Street
Third Floor
San Jose, California 95113

5 METHOD FOR ASSIGNING A STREAMING MEDIA SESSION
TO A SERVER IN FIXED AND
MOBILE STREAMING MEDIA SYSTEMS

10 TECHNICAL FIELD

The present claimed invention relates to the field of streaming media. More specifically, the present claimed invention relates to delivering streaming data to fixed clients and/or mobile clients using multiple description bitstreams and various forms of diversity.

BACKGROUND ART

Today's networks, such as the Internet, are primarily designed for delivering static, non-real-time data to fixed clients such as desktop computers and laptops. Developing a system that delivers real-time streaming media to mobile clients presents an even greater series of challenges due to the streaming nature of the data and the mobility of the user. These challenges are intensified when considering issues such as system scalability, which extends service to a larger number of users; and quality-of-service and fault tolerance, which provides mobile users with continuous, uninterrupted streaming media sessions. Furthermore, this robust, uninterrupted media delivery session must be delivered over best-effort networks, which provide best effort, but non-guaranteed levels of service. A solution that overcomes these many challenges requires innovation throughout the end-to-end system.

Currently, multimedia applications such as video and audio communication over the Internet or wireless links are hampered by the limited bandwidth and losses (e.g. packet loss or bit errors) that afflict these error-prone environments. These multimedia applications require high compression and high error resilience, however simultaneously achieving these qualities is difficult because these are largely conflicting requirements.

Two important characteristics of a media communication system

- render both of them useless. When the same information is transmitted twice in a wired network, the information typically proceeds along the same path in the network. Therefore, if that same path is congested or experiences an outage, then both duplicates of the information would be lost. To summarize, in an attempt to improve reliability the conventional approach is to duplicate information and resources. This conventional approach is inefficient because of the duplication, and is also ineffective because in many cases there can still exist single points of failure.
- Although portions of the above-listed discussion specifically mention the shortcomings of prior art approaches with respect to the streaming of video data for simplified presentation, such shortcomings are not limited solely to the streaming of video data. Instead, the problems of the prior art span various types of media including, but not limited to, audio-based data, speech-based data, image-based data, graphic data, web page-based data, and the like. Moreover, streaming media typically shares the property that the media streams must be delivered with a relative time constraint and thus share the notion of a stream.
- Thus, the need has arisen for a method and system for streaming media to fixed clients and/or mobile clients. A further need exists for a method and system for streaming media to fixed clients and/or mobile clients wherein the method and system provides increased reliability and efficiency over conventional systems.

DISCLOSURE OF THE INVENTION

The present invention provides a method and system for streaming media to fixed clients and/or mobile clients. The present invention further provides a method and system for streaming media to fixed clients and/or mobile clients wherein the method and system provides increased reliability and efficiency over conventional systems.

Specifically, a method for assigning servers to provide multiple description bitstreams to a base station (in a mobile client environment) or to a fixed client (in a mobile client environment) is provided. In one embodiment, the present invention, upon receiving a request from a mobile client to have media data streamed thereto, analyzes a plurality of servers to determine a first candidate server for providing a first multiple description bitstream to the base station along a first path. The present method also determines a second candidate server for providing a second multiple description bitstream to the base station along a second path. The present method then sends a request to the first candidate server to provide the first multiple description bitstream to the base station along the first path, and also sends a request to the second candidate server to provide the second multiple description bitstream to the base station along a second path. In another embodiment, there are two separate paths from two separate servers to two separate base stations and then from each base station there is a separate path to the mobile client. In still another embodiment, there are two paths from a single server to two separate base stations and then from each base station there is a separate path to the mobile client. In one fixed client embodiment, the present invention is able to assign a plurality of servers to provide a plurality of MD bitstreams to the fixed client.

These and other technical advantages of the present invention will no doubt become obvious to those of ordinary skill in the art after having read the following detailed description of the preferred embodiments which are illustrated in the various drawing figures.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and form a part of this specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention:

FIGURE 1 is a schematic diagram of an exemplary computer system used to perform steps of the present method in accordance with various embodiments of the present claimed invention.

FIGURE 2 is a schematic diagram illustrating multiple description coding of media data as employed in accordance with various embodiments of the present claimed invention.

FIGURE 3A is a schematic diagram illustrating a mobile client system employed in accordance with various embodiments of the present claimed invention.

FIGURE 3B is a schematic diagram illustrating a fixed client system employed in accordance with various embodiments of the present claimed invention.

FIGURE 4 is a flow chart of steps performed in accordance with one embodiment of the present claimed invention.

FIGURE 5 is a schematic diagram of a content server, a plurality of servers having MD bitstreams stored thereon, and a series of cells in accordance with various embodiments of the present claimed invention.

FIGURE 6 is a flow chart of steps performed in accordance with one embodiment of the present claimed invention.

FIGURE 7A is a schematic diagram of a plurality of servers operating in conjunction with a network management protocol, a series of base stations, and a plurality of mobile clients in accordance with various embodiments of the present claimed invention.

FIGURE 7B is a schematic diagram of a plurality of servers operating in conjunction with a network management protocol, a series of

base stations, and a plurality of mobile clients in accordance with various embodiments of the present claimed invention.

FIGURE 7C is a schematic diagram of a plurality of servers
 5 operating in conjunction with a network management protocol, a series of
 base stations, and a plurality of mobile clients in accordance with various
 embodiments of the present claimed invention.

FIGURE 8 is a flow chart of steps performed in accordance with
 10 one embodiment of the present claimed invention.

The drawings referred to in this description should be understood
 as not being drawn to scale except if specifically noted.

09800000 070301

BEST MODES FOR CARRYING OUT THE INVENTION

Reference will now be made in detail to the preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings. While the invention will be described in conjunction with the preferred embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the spirit and scope of the invention as defined by the appended claims. Furthermore, in the following detailed description of the present invention, numerous specific details are set forth in order to provide a thorough understanding of the present invention. However, it will be obvious to one of ordinary skill in the art that the present invention may be practiced without these specific details. In other instances, well known methods, procedures, components, and circuits have not been described in detail as not to unnecessarily obscure aspects of the present invention.

It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions utilizing terms such as "encoding", "transmitting", "storing", "distributing" or the like, refer to the actions and processes of a computer system, or similar electronic computing device. The computer system or similar electronic computing device manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission, or display devices. The present invention is also well suited to the use of other computer systems such as, for example, optical and mechanical computers.

COMPUTER SYSTEM ENVIRONMENT OF THE PRESENT INVENTION

With reference now to Figure 1, portions of the present method and system are comprised of computer-readable and computer-executable instructions which reside, for example, in computer-usable media of a computer system. Figure 1 illustrates an exemplary computer system used in accordance with one embodiment of the present invention. It is

appreciated that system 100 of Figure 1 is exemplary only and that the present invention can operate on or within a number of different computer systems including general purpose networked computer systems, embedded computer systems, routers, switches, server devices, client devices, various intermediate devices/nodes, stand alone computer systems, and the like. Additionally, computer system 100 of Figure 1 is well adapted having computer readable media such as, for example, a floppy disk, a compact disc, and the like coupled thereto. Such computer readable media is not shown coupled to computer system 100 in Figure 1 for purposes of clarity. Additionally, portions of the present embodiment are well suited to operating in conjunction with various mobile clients such as, for example, a cell phone, personal digital assistant (PDA), laptop computer, pager, and the like.

System 100 of Figure 1 includes an address/data bus 102 for communicating information, and a central processor unit 104 coupled to bus 102 for processing information and instructions. Central processor unit 104 may be an 80x86-family microprocessor. System 100 also includes data storage features such as a computer usable volatile memory 106, e.g. random access memory (RAM), coupled to bus 102 for storing information and instructions for central processor unit 104, computer usable non-volatile memory 108, e.g. read only memory (ROM), coupled to bus 102 for storing static information and instructions for the central processor unit 104, and a data storage unit 110 (e.g., a magnetic or optical disk and disk drive) coupled to bus 102 for storing information and instructions. System 100 of the present invention also includes an optional alphanumeric input device 112 including alphanumeric and function keys coupled to bus 102 for communicating information and command selections to central processor unit 104. System 100 also optionally includes an optional cursor control device 114 coupled to bus 102 for communicating user input information and command selections to central processor unit 104. System 100 of the present embodiment also includes an optional display device 116 coupled to bus 102 for displaying information.

Referring still to Figure 1, optional display device 116 of Figure 1, may be a liquid crystal device, cathode ray tube, or other display device suitable for creating graphic images and alphanumeric characters recognizable to a user. Optional cursor control device 114 allows the computer user to dynamically signal the two dimensional movement of a

visible symbol (cursor) on a display screen of display device 116. Many implementations of cursor control device 114 are known in the art including a trackball, mouse, touch pad, joystick or special keys on alphanumeric input device 112 capable of signaling movement of a given direction or manner of displacement. Alternatively, it will be appreciated that a cursor can be directed and/or activated via input from alphanumeric input device 112 using special keys and key sequence commands. The present invention is also well suited to directing a cursor by other means such as, for example, voice commands. A more detailed discussion of the present invention is found below.

GENERAL METHOD AND SYSTEM FOR RELIABLY AND EFFICIENTLY STREAMING MEDIA TO FIXED AND/OR MOBILE CLIENTS

As an overview, the present invention provides a method and system for streaming media to fixed and/or mobile clients wherein the method and system achieve improved efficiency by using complementary information (as opposed to duplicate information employed in conventional approaches). The present invention further provides a method and system for streaming media to fixed and/or mobile clients wherein the method and system improve reliability by using various forms of diversity to prevent single points of failure.

For purposes of clarity and brevity, the following discussion and examples will specifically deal with video data. The present invention, however, is not limited solely to use with video data. Instead, the present invention is well suited to use with audio-based data, speech-based data, image-based data, web page-based data, graphic data and the like. Of course, buffering and downloading are allowed, and various degrees of buffering can be accommodated in this system; thus, this system supports both live and buffered media streams. Furthermore, note that all these media types can be coded with the properties of multiple description bitstreams described herein.

One approach for reliably delivering streaming media is disclosed in commonly-owned, co-pending U.S. Patent Application Serial No. 09/400,416, entitled "Video Communication Using Multiple Streams", filed September 21, 1999 to J. G. Apostolopoulos. Another approach for reliably delivering streaming media is disclosed in commonly-owned, co-pending U.S. Patent Application Serial No. 09/784,226, entitled "Method and

5 Communication System Employing Multiple State Encoding and Path
Diversity", filed January 19, 2001 to J. G. Apostolopoulos. U.S. Patent
Application Serial No. 09/400,416, U.S. Patent Application Serial No.
09/784,226, and U.S. Patent Application Serial No. 09/784,223 are each
incorporated herein by reference as background material. These patent
10 applications relate to a system for reliable video communication over lossy
packet networks while preserving high compression efficiency. Portions
of the prior work were composed of two systems: (1) a multiple description
video coding system, and (2) a path diversity transmission system.

With reference now to Figure 2, a schematic diagram illustrating multiple description coding of media data as employed in conjunction with various embodiments of the present invention is shown. Multiple Description Coding (MDC) refers to one form of compression where the goal is to code an incoming signal into a number of separate bitstreams, where the multiple bitstreams are often referred to as multiple descriptions. These separate bitstreams have the property that they are all independently decodable from one another. Specifically if a decoder receives any single bitstream it can decode that bitstream to produce a useful signal (without requiring access to any of the other bitstreams). MDC has the additional property that the quality of the decoded signal improves as more bitstreams are accurately received. For example, assume that a video is coded with MDC into a total of N streams. As long as a decoder receives any one of these N streams it can decode a useful version of the video. If the decoder receives two streams it can decode an improved version of the video as compared to the case of only receiving one of the streams. This improvement in quality continues until the receiver receives all N of the streams, in which case it can reconstruct the maximum quality. As shown in Figure 2, a multiple description encoder 202 codes an original signal into two streams, referred to as stream 1 and stream 2. In the example of Figure 2, there are three decoders 204, 206, and 208. Each of decoders 204, 206, and 208 receive different bitstreams. Decoder 1 204 receives only stream 1 and decodes that stream to produce usable video. Decoder 2 206 receives only stream 2 and decodes that stream to produce usable video. Decoder 3 208 receives both stream 1 and

There is a critical difference between the present multiple description based approach for streaming media delivery and prior approaches, such as scalable or layered coding approaches for streaming media delivery. Namely, in scalable or layered coding the video is also coded into multiple bitstreams, however one bitstream, referred to as the baselayer bitstream, is critically important and must be correctly received in order to produce a usable decoded media stream. Specifically, in the conventional scalable or layered approaches for streaming media delivery, even if all the bitstreams except the baselayer bitstream are correctly received, they are essentially useless unless the baselayer bitstream is correctly received, creating a single point of failure. The present multiple description based streaming media delivery does not have this problem since as long as any multiple description bitstream is received it can be decoded to produce usable quality video, and as more multiple description bitstreams are received the quality of the decoded video increases.

A preferred embodiment of MDC video coding that provides the above properties and achieves high compression is given in the prior patent. This MDC video coding system does not require a back-channel and therefore can be applied in a wide variety of applications (e.g. broadcast or multicast). In addition, it has the attractive property that it can be applied as a standard-compatible enhancement within MPEG-4

Version 2 (with NEWPRED) and H.263 Version 2 (with RPS). Therefore any MPEG-4 Version 2 decoder can decode the resulting bitstream while an enhanced decoder designed to perform state recovery as presented in the prior patent can provide improved error recovery. This preferred
 5 embodiment of MDC video coding is assumed through the following discussion. However, a different video compression algorithm that has the same MDC properties as discussed above may also be used in its place.

Following is a discussion of path diversity as employed in
 10 conjunction with various embodiments of the present invention. Consider the case of multimedia communication over a packet network such as the Internet. Communication over the Internet is often hampered by congestion and packet loss. An important observation is that while one node or path in the network may be congested, other nodes or paths may
 15 have ample bandwidth. It would be advantageous to know the instantaneous quality of each path and to use that information to send packets along the "best" path (much like listening to a traffic report before leaving for work). However this is very difficult for a number of reasons, including the fact that the congested areas can vary quite rapidly.

20 While it may not be possible to know which paths are the best to use at any point in time, through appropriate system design one can still achieve significant performance improvements. Various embodiments of the present invention employ a path diversity system which explicitly
 25 sends different subsets of packets for an application over different paths, as opposed to the default scenario where the stream of packets proceeds along a single path. By using multiple paths at the same time some amount of averaging occurs and the end-to-end application effectively sees an "average" path behavior. Generally, seeing this average path behavior
 30 provides better performance than seeing the behavior of any randomly chosen individual path. For example, the probability that all of the multiple streams that are transmitted on different paths are simultaneously congested and have losses is much less than the probability that a single path is congested. The benefits of path diversity
 35 include (1) the application sees a virtual average path which exhibits a smaller variability in communication quality than exists over an individual path, (2) burst packet losses are converted to isolated packet losses, and (3) the probability of an outage (where all packets in the packet stream are lost for the duration of the outage) is greatly reduced. These

advantages provide some important benefits to multimedia communication performance under packet loss. As will be discussed in detail below, the various embodiments of the present invention routes MDC traffic through semi-intelligent nodes at strategic locations in the Internet, thereby providing a service of improved reliability while leveraging the infrastructure of the Internet.

Path diversity may also exist in wireless networks. Various embodiments of the present invention employ a soft-handoff system in which a mobile client can simultaneously communicate with multiple base stations. In such cases, the benefits of path diversity mentioned above are also realized in a wireless environment.

Additionally, the MD video coding and path diversity employed in conjunction with the various embodiments of the present invention are useful even if used separately. For example, MD video coding can provide improved reliability even when sent over a single path. Similarly, path diversity provides a virtual channel with improved characteristics, enabling a simpler system design. However, when used together, MD video coding and path diversity complement, and also to a certain extent enhance, each other's capabilities. MD video coding provides multiple independently decodable bitstreams, which the transmission system explicitly sends over different paths, and the transmission system provides the video decoder with a high probability that at least one of the streams will be received correctly at any point in time. In one preferred embodiment of MD video coding, this enables the video decoder to perform state recovery to recover a corrupted stream.

With reference next to Figure 3A and flow chart 400 of Figure 4, exemplary steps used by the various embodiments of present invention are illustrated. Flow chart 400 includes processes of the present invention which, in one embodiment, are carried out by a processor under the control of computer-readable and computer-executable instructions. The computer-readable and computer-executable instructions reside, for example, in data storage features such as computer usable volatile memory 106, computer usable non-volatile memory 108, and/or data storage device 110 of Figure 1. The computer-readable and computer-executable instructions are used to control or operate in conjunction with, for example, central processing unit 104 of Figure 1.

With reference again to Figure 3A, a system 300 that delivers streaming media to mobile clients over hybrid wired/wireless networks in accordance with one embodiment of the present invention is shown. In one embodiment, system 300 consists of one or more servers (304a-304e), one or more wireless base stations (306a and 306b), and one or more mobile clients (e.g. cellphone 302 and/or personal digital assistants (PDAs) 308a and 308b) as shown in Figure 3A. The system of the present invention may include a greater or lesser number of components than are specifically illustrated in the embodiment of Figure 3A. As an example, although not always required, a content server 310 also forms a portion of the system of the embodiment of Figure 3A. Importantly, in the following discussion, the term "server" in various embodiments is intended to encompass a device functionally resembling a computer (e.g. having computation ability, memory, and/or connectivity capability). A typical server according to the definition as used in the present application may include, but is not limited to, any computer (e.g. mainframe, corporate server, personal computer (PC), laptop, personal digital assistant (PDA), and the like). In various other embodiments of the present invention, the term "server" is intended to encompass a device not typically considered a computer but having similar capabilities. In such an embodiment, the server is comprised, for example, of an advanced cell phone.

Importantly, it should be noted that the methods of various embodiments of the present invention are applicable with fixed wired clients and/or mobile wireless clients. Specifically, the mobile client case is a more general and superset version of the fixed client case. For example, in the mobile client case, the MD bitstreams are provided by a server or servers to a mobile client through one or more base stations. In contrast, the corresponding fixed (wired) client case would have the server or servers instead provide the MD bitstreams directly to the fixed client without the need for a base station. Therefore, in the following discussion, will specifically discuss the more general and superset mobile client case. For purposes of brevity and clarity, redundant examples of fixed client cases are not presented herein. It will be understood by one of ordinary skill in the art, however, that in an example in which MD bitstreams are provided by a server or servers to a mobile client by one or more base stations, in the fixed client case the server or servers would instead provide the MD bitstreams to the fixed client without the need for a base

station. With reference now to Figure 3B, a system 350 that delivers streaming media to fixed clients over a network (e.g. the Internet) in accordance with one embodiment of the present invention is shown. In one embodiment, system 350 consists of a one or more fixed clients (e.g. 5 personal computers 352 and 356), one or more servers (304a-304e), a content server 310. In the present embodiment fixed client 352 is coupled to network 351 via a wired link 354. Similarly, fixed client 356 is coupled to network 351 via a wired link 358. The system of the present invention may include a greater or lesser number of components than are specifically 10 illustrated in the embodiment of Figure 3B. As an example, although not always required, content server 310 also forms a portion of the system of the embodiment of Figure 3B. As mentioned above, in the following discussion, the term "server" is in various embodiments intended to encompass a device functionally resembling a computer (e.g. having 15 computation ability, memory, and/or connectivity capability). A typical server according to the definition as used in the present application may include, but is not limited to, any computer (e.g. mainframe, corporate server, personal computer (PC), laptop, personal digital assistant (PDA), and the like). In various other embodiments of the present invention, the 20 term "server" is intended to encompass a device not typically considered a computer but having similar capabilities. In such an embodiment, the server is comprised, for example, of an advanced cell phone.

Furthermore, it should be noted the present invention is well suited 25 to use in any of a variety of wired and/or wireless networks, and also in the various combinations thereof. For example, embodiments of the present invention are well suited to use in: wired and wired networks (e.g. a wired network infrastructure also having a wired connection to the client); wired and wireless networks (e.g. a wired network infrastructure 30 having a wireless connection to the client); wireless and wired networks (e.g. a wireless network infrastructure having a wired connection to the client); and wireless and wireless networks (e.g. a wireless network infrastructure also having a wireless connection to the client). More specifically, the methods of the present invention are also applicable if all 35 of the links are wireless, as in the case of an ad-hoc wireless network where some of the nodes are acting as servers and others as clients. Note that a node may act as both a server (source of information) and a client (sink of information) at the same time. An interesting example of this later case is when each node is a laptop or PDA with a wireless

connection. Also note that each cell phone may act as both a sender and receiver of MD streams in a wireless/wired/wireless situation.

Referring now to Figures 3A and 3B, in following discussions in which, for example, MD bitstreams are provided by a server or servers to a mobile client through one or more base stations (as shown in Figure 3A), the corresponding fixed client case would have the server or servers instead provide the MD bitstreams to the fixed client without the need for a base station. Hence, it can be seen that the below recited mobile client-based examples and discussions are applicable as well to fixed client systems. That is, the following mobile client-based examples and discussions are not intended to limit the present invention to applicability only in mobile client systems.

With reference now to flow chart 400 of Figure 4, steps performed in accordance with one embodiment of the present invention are shown. Although specific steps are disclosed in flow chart 400 of Figure 4, such steps are exemplary. That is, the present invention is well suited to performing various other steps or variations of the steps recited in Figure 4. At step 402, the present embodiment encodes data to be streamed to a mobile client into two or more MD streams. In one embodiment, the data to be streamed is comprised of a video sequence. As mentioned above, MD streams have the property that any subset of them can be decoded into a media stream whose quality depends on the number of decoded streams. In the present embodiment, this encoding may be done in real time or it may be done in advance in which case the pre-computed MD streams are stored on a content server (e.g. content server 310 of Figure 3).

Referring still to step 402, in the present embodiment, the present invention uses specially designed multiple description media streams which contain complementary information (as opposed to duplicating the information). Specifically, the original media stream is coded using a multiple description algorithm into a number of separate descriptions or bitstreams. These descriptions have the property that (1) each bitstream is independently useful to the client, and (2) each bitstream contains complementary information. For example, consider the case of multiple description coding with two descriptions. As long as the receiver receives either of the bitstreams, it can decode a usable media stream. If the receiver receives both bitstreams it can decode a higher quality media

stream than if it had received either bitstream alone. In addition, the MD streams provide these properties while requiring only slightly higher total bit rate than that required by a conventional coding algorithm that does not provide these properties. Additionally, an important point is that each description or MD bitstream is equally important. This is in contrast to conventional scalable schemes, where the base-layer bitstream is critically important. That is, in conventional scalable schemes, if the base-layer bitstream is lost the other bitstream(s) are useless. Specifically, since each MD bitstream is equally important in the present embodiment, there is no single point of failure in the sense that there is no single bitstream that must be received.

At step 404, the present embodiment then distributes the MD streams to a number of different servers (e.g. servers 304a-304e of Figure 3) placed at intermediate nodes throughout a network. By, appropriately distributing the MD streams, the present invention eliminates the possibility that any single fault may render all streams useless. In one embodiment, the present invention uses servers that are placed at intermediate nodes in the network, for example alongside a router or a wired/wireless gateway. In the present embodiment, servers 304a-304e send the MD streams to nearby wireless base stations 306a and 306b as mobile clients 302, 308a, and 308b roam through their coverage areas. Wireless base stations 306a and 306b receive data from the wired network and wirelessly transmit this data to mobile clients 302, 308a and 308b. Likewise, wireless base stations 306a and 306b wirelessly receive data from mobile clients 302, 308a and 308b and transmit this data to the wired network. Thus, wireless base stations 306a and 306b can be viewed as having a wired/wireless gateway and a wireless transmitter/receiver. Furthermore, as will be described in detail below, the various embodiment of the present invention overcome the non-guaranteed, best-effort nature of existing networks by dynamically delivering MD streams to mobile users from the most accessible servers based on user mobility, network congestion, and server load.

With reference still to step 404, the present invention ensures that no single fault causes the loss of all streams. In the context of the above example, a video sequence media stream may be coded into two MD streams which are then placed at two different storage modules, connected to two different servers at two different locations in the network.

In so doing, the two MD streams can be sent over two different paths in the wired network, and two different base stations may transmit the two MD streams over two different wireless channels to the receiving client. In this manner, the present embodiment achieves a level of diversity

5 which eliminates many potential single points of failure.

Referring now to step 406, the present embodiment provides access to the MD coded streams for a mobile client. That is, the MD streams are accessible to, for example, a requesting cell phone, PDA, laptop computer, or other such mobile client.

Referring again to steps 402, 404 and 406, a detailed example (using Figure 3 for illustration) is provided below of the operation of the present embodiment. In the present example, a video sequence media stream has previously been coded into two MD streams which were then both stored on content server 310. The first of the two MD streams is placed at a first storage module, coupled to server 304a. The second of the two MD streams is placed at a second storage module coupled to server 304e. When the video sequence media stream is requested, for example, by cell phone 302, the first of the two MD streams is sent over a first path to cell phone 302, and the second of the two MD streams is sent over a second path to cell phone 302. In the present example, the first path comprises transmission over a wired network connection from server 304a ultimately to wireless base station 306a, and then wirelessly transmitting the first MD stream of the requested video sequence from wireless base station 306a to cell phone 302. Furthermore, in the present example, the second path comprises transmission over a wired network connection from server 304e ultimately to wireless base station 306b, and then wirelessly transmitting the second MD stream of the requested video sequence from wireless base station 306b to cell phone 302. The MD streams of the present embodiment have the property that any number of streams can be decoded into a sequence in which the quality of the decoded sequence depends on the number of decoded MD streams. Specifically, any one MD stream can be decoded into baseline quality data; any two MD streams can be decoded into improved quality data; and so on until finally all the MD streams can be decoded into the highest quality data. Thus, should one of the paths fail in the preceding example (e.g. the second path), cell phone 302 is still able to receive and utilize the requested video sequence based solely on receiving only one of the MD streams.

Although such first and second paths are recited in the present example, the present invention is well suited to use with the numerous paths which can be constructed within, for example, the network of system 300. Additionally, although two MD streams are recited in the present description, the present invention is well suited to use with more than two MD streams of data. Thus, the present embodiment, composed of multiple description coding and system diversity, provides improved system reliability in a more efficient manner than the conventional approach of resource duplication. Furthermore, the proposed system enables the elimination of a number of single points of failure that afflict conventional systems. Also, as mentioned above, for purposes of clarity and brevity, the above and following discussion and examples may specifically deal with video data. The present invention, however, is not limited solely to use with video data. Instead, the present invention is well suited to use with audio-based data, speech-based data, image-based data, graphic data, web page-based data, and the like.

Thus, the present invention provides a method and system for streaming media to fixed clients and/or mobile clients. The present invention further provides a method and system for streaming media to fixed clients and/or mobile clients wherein the method and system provides increased reliability and efficiency over conventional systems.

Furthermore, the multiple description bitstreams may be placed on servers in a variety of ways depending on the specific situation. For example, some servers may store all of the descriptions, while other servers may store only a subset of the descriptions. An example of the former is a central server which has high connectivity to a large number of clients, it may store all of the descriptions and adaptively choose to transmit a specific subset of the descriptions to each specific client based on the client's particular situation (e.g. the other servers that the client is connected to and the descriptions that these other servers can provide). On the other hand there may be other servers for which only a subset of the descriptions are stored, for example only the first description may be stored on some servers and the second description may be stored on other servers. The strategy for distributing descriptions to different servers may depend on a number of factors such as each server's compute and storage capabilities, its connectivity, typical network conditions, disjointness of

paths, and the popularity of the specific media to be delivered.

When coding a specific media sequence, the media may be coded into multiple descriptions where each description requires the same bit rate and provides approximately the same quality. This may be referred to as balanced multiple description coding. Alternatively, the media may be coded into multiple descriptions where each description may require a different bit rate and may provide a different quality. This may be referred to as unbalanced multiple description coding. Coding a media into unbalanced multiple descriptions is important in situations where unbalanced operation is required, e.g. when one has unbalanced storage available at different servers or unbalanced available bandwidths on different paths. For both balanced and unbalanced multiple description coding, the key property is that as long as the client receives any subset of the multiple descriptions it can produce a usable decoded media stream, and that as it receives more descriptions the quality of the decoded media stream would increase.

A media sequence may be initially coded into balanced multiple description bitstreams, where each bitstream requires approximately the same bit rate, and these multiple descriptions streams may be appropriately placed at different servers in the network. These MD bitstreams may be subsequently transcoded to different lower bit rates in order to appropriately match the bandwidths available to a client at any particular time. For example, if a specific client requires unbalanced operation, e.g. if there are two paths available to the client and the available bandwidth of each path is different, the servers can transcode the multiple description bitstreams to the appropriate bit rates available on each path. Furthermore, the available bit rate along a specific path may vary as a function of time, and the server responsible for delivering the multiple description bitstream along that path can then appropriately transcode the bitstream to match the available bit rate on that path as a function of time. In this manner, the system can efficiently use the available bandwidth to maximize the reconstructed quality at the client.

The distribution of multiple description bitstreams to different servers, as well as the assignment of different servers to transmit different MD bitstreams to a specific client, depends on a number of factors including: demand on each server, available bandwidth and loss along the path from each server to the client, and the disjointness of the

multiple paths.

The issue of disjointness does not arise in prior work since conventional approaches transmit a single bitstream along a single path.

5 In the proposed approach multiple bitstreams are explicitly transmitted along multiple paths, and it is desired that these multiple paths be as disjoint as possible in order to minimize the probability that a single fault may lead to the loss of all of the multiple descriptions. For example, in the case of two MD bitstreams, each stream is sent over a separate path from
10 a server to the client. Each of these paths consists of a sequence of links, and these two paths may include a number of shared links as well as a number of links that are not shared. Shared links are referred to as joint links and unshared links are referred to as disjoint links. Ideally, all of the links on the two paths are disjoint. However, in practice this may
15 sometimes not be possible. The primary goal is to minimize the number of lossy joint links. For example, if an approximately lossless backbone link is joint or shared by both paths it will not effect the communication quality since the link is lossless. However, a lossy joint link can have a detrimental effect on the performance of the system since any losses, e.g.
20 produced by congestion, may lead to the loss of both descriptions. Therefore, given a particular client, the system carefully chooses two servers to send the two multiple descriptions to the client such that these servers have maximally disjoint paths between them and the client. Furthermore, to distribute MD streams on a set of servers, the anticipated
25 location of clients as well as the connectivity between each server and the anticipated clients are taken into account in order to determine the distribution that would enable the maximally disjoint paths between servers and anticipated clients. Of course, other more conventional metrics such as available bandwidth and losses on each link are also
30 taken into account in this optimization. These conventional metrics are not discussed as they are also used in conventional systems.

INITIAL MULTIPLE DESCRIPTION BITSTREAM PLACEMENT AND REDISTRIBUTION ON SERVERS

35 As will be discussed in greater detail below, the MD media bitstream system of the various embodiments of the present invention must perform a number of steps prior to and during a streaming session. The present section specifically discusses the assignment and distribution to servers, e.g. servers 304a-304e, located throughout a network such as is
40 shown in Figure 3.

In the present embodiment, prior to or upon a client request, the appropriate data (e.g. a video sequence) is coded into MD bitstreams. Next, the MD bitstreams are assigned and distributed to servers that are located throughout the network. To dynamically minimize end-to-end (i.e. server-to-client) delay, content is placed at servers at regions with highly anticipated demand (e.g. hot spots). Moreover, MD bitstreams are distributed among servers in a way so any client can find a complete set of MD bitstreams in its closest vicinity.

The present embodiment provides a method to 1) initially populate servers with MD bitstreams from content providers, and 2) redistribute the MD bitstreams among servers after the initial population. The initial server population method of the present embodiment takes place when fresh content needs to be distributed initially to servers in anticipation of client requests, and also when a client requesting content encounters a miss at one or more servers in the client's vicinity. As an example, the present initial server population method is employed when a client requests a movie trailer that has not yet been loaded to any server.

The redistribution of the MD bitstreams among the servers method of the present embodiment takes place when parameters such as, for example, usage patterns, mobility trends, network conditions, disjointness of paths to mobile clients, computation sizes and/or bitstream sizes change. One embodiment of the present method also considers such parameters during initial population of the servers with the MD bitstreams. Other parameters which provoke the redistribution of the MD bitstreams among the servers include, for example, network connectivity, network topology, server load, server storage space, server availability, and server I/O bandwidth. The present initial server population method is employed, for example, when, as a client moves, MD bitstreams are pre-fetched into servers at the new region if not available already.

The methods of the present embodiment will now be described in conjunction with Figure 5 and flow chart 600 of Figure 6. Figure 5 is a schematic diagram illustrating a content source 502, a plurality of servers 504a-504j, and a plurality of cells 506 including cell 506a and cell 506b. Although specific steps are disclosed in flow chart 600 of Figure 6, such steps are exemplary. That is, the present invention is well suited to

performing various other steps or variations of the steps recited in Figure 6. Furthermore, the following discussion will begin with the initial server population method of the present embodiment. With reference now to step 602 of Figure 6, the present method first identifies the cell, c_0 , with the

highest demand estimated in real-time, or anticipated highest demand in the foreseeable future above a certain threshold value, v , i.e. the present embodiment identifies the "hottest spot". If no such cell exists, then there is not enough client concentration to require server population according to the present embodiment. If such a cell does exist, the present

embodiment proceeds to step 604. The present invention is well suited to varying the threshold value, v , to a desired value. For purposes of illustration, it will be assumed that cell 506a of Figure 5 is found to be the cell with the highest demand above the threshold value, v . Also, in another embodiment of the present invention, a server is populated with the MD bitstreams when it exhibits high connectivity in the network, and, thus, is able to readily serve nearby hot spots with low delay and high bandwidth. In other words, in one such embodiment, even if a server is not located at the "hottest spot" or even at a hot spot, if it is close to a hot spot in terms of network conditions, then it can be populated with MD bitstreams.

At step 604, the present embodiment then initializes the number of servers (S) required to service this "hottest" cell, d , to be N , where N is the number of MD bitstreams composing the content to be streamed to the mobile client. Again for purposes of illustration, in the present embodiment, a video sequence has been coded into two MD bitstreams. The first MD bitstream is represented as "0" and the second MD bitstream is represented by a "1". Although only two MD bitstreams are recited in the above embodiment, the present invention is well suited to media coded into more than two MD bitstreams. Additionally, for purposes of clarity and brevity, the above and following discussion and examples may specifically deal with a video sequence. The present invention, however, is not limited solely to use with a video sequence. Instead, the present invention is well suited to use with audio-based data, speech-based data, image-based data, graphic data, web page-based data, and the like. Also, although d is set to be equal to N in the present example, the present invention is also well suited to setting the value, d , to be other than equal to N .

With reference still to step 604, because two MD bitstreams are recited above (i.e. because N is 2 in the present example), the present embodiment sets the number of servers (S), d, required to service the "hottest cell, cell 506a, at two.

5

At step 606, the present embodiment then finds the d servers closest to the cell, c0, and determines whether those servers contain the N MD bitstreams. That is, the present method determines which subset of the available servers are disposed close to cell, c0. Thus, in the present example, the present embodiment finds the two servers which are closest to cell 506a. Note that in the present embodiment, "closeness" is weighted by the current network conditions so that a server with congested links will appear "far" from the hottest cell, c0, even though it may be topological close. In the present example, servers 504b and 504d are deemed closest to cell 506a. In the present example, the present method then determines if the two MD bitstreams are present at servers 504b and 504d. Furthermore, the present embodiments are metric-independent and, hence, will work for other convenient metrics of closeness such a geographical distance.

10

15

20

With reference still to step 606, if the closest servers contain the N MD bitstreams, then the present method is finished with the hottest cell, c0, and moves on to the next hottest cell (i.e. return to step 602). If the closest servers do not contain the N MD bitstreams, the present embodiment moves on to perform step 608. Thus, in the present example, if the two MD bitstreams of the video sequence are present at servers 504b and 504d, then the present method returns to step 602. If, on the other hand, the two MD bitstreams of the video sequence are not present at servers 504b and 504d, then the present method proceeds to step 608.

25

30

35

Thus, in the illustration of Figure 5, because the first MD bitstream, 0, is present at server 504b, and the second MD bitstream, 1, is present at server 504d, the present method would return to step 602. If, however, either the first MD bitstream, 0, or the second MD bitstream, 1, was not present at any of servers 504b and 504d, the present method would proceed to step 608. Such a condition would occur, for example, during an initial population of the servers according to the present method. That is, upon identification of the hottest cell, the required number of servers, and the determination of the closest servers (and assuming no prior population of the current MD bitstreams), it would be found that the MD bitstreams of

interest would not be present at the closest servers (or any other servers). Thus, barring any prior population of the MD bitstreams of interest, the present method will ordinarily proceed to step 608.

5 At step 608, the present embodiment then determines if it is effective to distribute all N MD bitstreams to the d servers. As an illustrative example, it may not be effective to distribute all N MD bitstreams to the d servers when one or more of the servers are already very heavily loaded, lacks sufficient memory to store an MD bitstream, or is otherwise
10 unsuited for the desired storage of the MD bitstream. In one embodiment, the present distribution step is dependent on server-dependent conditions such as load. For example, an MD bitstream is not placed on a particular server if that server is loaded over some specified threshold. The present invention is well suited to altering the aforementioned specific threshold
15 to accommodate various objectives. As a result, in the present embodiment, a server that is lightly loaded is chosen over one that is closer to the hot spot (i.e. hottest cell, c0) but which already carries a heavy load. In the present example, if it is assumed that, for example, either the first MD bitstream, 0, or the second MD bitstream, 1, was not present at
20 any of servers 504b and 504d, the present method would then determine if the two MD bitstreams could be distributed to servers 504b and 504d. If it is effective to distribute all N MD bitstreams to the d servers, the present method proceeds to step 612 described below. If, on the other hand, it is not effective to distribute all N MD bitstreams to the d servers, then the
25 present method proceeds to step 610.

At step 610, if it was determined at step 608 that it is not effective to distribute all N MD bitstreams to the d servers, the present method increments the number of servers, d. In so doing, the present method
30 increases the potentially available server space and then returns to step 606. Depending on the implementation, step 606 and step 608 do not have to consider the servers that were deemed not suitable for MD bitstream placement in the next iteration. That is, in one embodiment, the present invention increases the search space of suitable MD bitstream placement,
35 but it is not necessary to consider the servers that did not pass step 606 and/or step 608 previously, unless the load conditions on those servers changed in the meantime. With reference to the present example, if the two servers 504b and 504d were not able to accommodate the storage of the two MD bitstreams, the number of servers would be increased from two to

three. In one embodiment, the number of servers is incremented by selecting the next closest server to the hottest cell, c0. The present example would then return to step 606 and continue from that point.

5 At step 612, if it was determined at step 608 that it is effective to distribute all N MD bitstreams to the d servers, the present method then distributes the N MD bitstreams to the d servers. Hence, with reference to the present example, if the two servers 504b and 504d were cumulatively able to accommodate the storage of the two MD bitstreams thereon (and
10 two MD bitstreams had not previously been stored thereon), the two MD bitstreams would be stored cumulatively on servers 504b and 504d. As a more specific example, in one embodiment, at least one of the MD bitstreams is stored on server 504b (e.g. MD bitstream 0) and at least one of the MD bitstreams (e.g. remaining MD bitstream 1) is stored on server
15 504d.

While the above description refers primarily to the first application of the present method (i.e. initial population of servers with MD bitstreams from content providers), the second portion of the present
20 method (i.e. redistribution of the MD bitstreams among servers after the initial population) is somewhat similar. Specifically, in the case of redistribution of the MD bitstreams among servers after the initial population, the present method begins the method of steps 602-612 with updated user statistics to identify the hottest cell, and to determine if the
25 new hot spots (i.e. hottest cells) are being properly serviced. Hence, the present embodiment enables dynamic reassigning of MD bitstreams based upon user defined criteria. As an example, in one embodiment, the method of the present invention (i.e. steps 602-612) is restarted each time interval of a predefined duration. In another embodiment, the method of
30 the present invention is restarted based upon a hit or miss rate at a given server. In still another embodiment, the method of the present invention is restarted based upon certain known traffic patterns to which the network is subjected. In yet another embodiment, the method of the present invention is restarted based upon network congestion conditions.
35 Various other embodiments of the present invention reassign MD bitstreams based upon other criteria. For example, in various embodiments, MD bitstreams are reassigned when: a server is overloaded (i.e. the server is running out of computational cycles); the storage capacity of at a server is exceeded; or upon network partition (e.g. link

failure which is different from congestion). Although such specific examples are recited above, the present method is well suited to dynamically reassigning the MD bitstreams to servers after the initial population thereof based upon various other parameters.

5

The present embodiment is also well suited to varying the method recited in steps 602-612 of Figure 6 such that the data to be streamed is initially populated on or redistributed to portions of a network where greater activity is expected. For example, in one embodiment of present invention, the MD bitstreams are stored on or redistributed to those servers which are proximate to or accessed from a heavily traveled commute corridor. In so doing, the MD bitstreams are disposed more closely to the large quantity of mobile clients which are expected to be traveling along commute corridor. In one embodiment of the present invention, during non-commute hours, the MD bitstreams are then redistributed from those servers which are proximate to or accessed from the heavily traveled commute corridor to a more appropriate location. More generally, the present method is well suited to varying the location at which the MD bitstreams are stored to accommodate anticipated mobile client location and demand.

10

15

20

Several distinct advantages are realized by the present method. For example, in conventional schemes, in order to have content to be streamed wholly available on a plurality servers, the entire content would have to be completely duplicated and placed in its entirety on every one of the plurality of servers. Such duplication is extremely inefficient and consumes valuable memory space as well as disk space. To the contrary, the present invention generates a plurality of MD bitstreams of the content to be streamed. As stated above, separate MD bitstreams have the property that they are all independently decodable from one another. Specifically if a decoder receives any single bitstream it can decode that bitstream to produce a useful signal (without requiring access to any of the other bitstreams). Thus, in the present embodiment, the number of generated bitstreams can be spread over a larger number of servers without requiring the inefficient duplication associated with conventional systems. As an example, assume that a video sequence is MD coded into four MD bitstreams. In the present method, the four MD bitstreams could be individually placed on four respective servers located throughout a network (i.e. one description to each of the four servers). Thus, the

25

30

35

present invention allows the data to be spread over a larger area thereby potentially providing easier access to a greater number of mobile clients, and/or increasing the range over which the data is readily available to a mobile client. Furthermore, the four MD bitstreams will cumulatively
 5 consume only slightly more memory than is necessary to store a single conventionally stored copy of the content to be streamed. In order to achieve such coverage using conventional methods, four complete duplicates of the entire content to be streamed would have to be placed on each of the above-mentioned four servers.

10 As yet another benefit, the present embodiment provides an intrinsic path diversity advantage without requiring inefficient complete duplication of data to be streamed. That is, using the above example of four MD bitstreams, the mobile client has the potential of requesting and
 15 receiving all four streams via four separate transmission paths. Thus, a single point of failure (e.g. a single server failure) or a network link failure (e.g. partitioning of the topology) does not prevent receipt by the mobile client of the streamed data.

20 As still another benefit, the present embodiment is also well suited to varying the method recited in steps 602-612 such that enhanced reliability is achieved. For example, a content provider may wish to ensure that mobile clients have superior access to the data to be streamed. In such a case, the content provider can opt to expand path diversity
 25 options and MD bitstream availability. In one extreme example, the method of the present embodiment would store all of the plurality of MD bitstreams on every available server such that the data is readily available to a mobile client in communication with the network. The present method is also well suited to a less drastic approach that does not store
 30 each of the plurality of MD bitstreams on every available server, but still provides enhanced reliability by increasing the number of servers on which the MD bitstreams are stored and/or by storing more than one of the MD bitstreams on the selected servers. More generally, the present method is well suited to varying the density of the stored MD bitstreams in
 35 the network.

As yet another benefit, the present invention is also well suited to transmitting MD bitstreams over a path which operates most effectively. That is, the present embodiment is also well suited to, for example,

transmitting all MD bitstreams from a single server (on which all MD bitstreams happen to be commonly stored) if such an approach proves most expeditious. More generally, the present method is not limited to, for example, transmitting each of the separate MD bitstreams to a mobile client via different respective paths.

As still another benefit, the present invention is also reduces input/output (I/O) bandwidth used on a server as compared to conventional approaches. That is, in the present embodiments, each MD bitstream is encoded at a lower bitrate than the original complete stream. Hence, the transmission of the MD bitstream can be accomplished at a bitrate which is lower than that required to transmit the original complete stream.

DYNAMIC SERVER SESSION ASSIGNMENT BASED ON SERVER AVAILABILITY, STREAM LOCATIONS, AND TRAFFIC CONDITIONS

As will be discussed in greater detail below, the MD media stream system of the various embodiments of the present invention must perform a number of steps prior to and during a streaming session. The present section specifically discusses the assignment of servers (e.g. servers 702a-704d) to a mobile client (e.g. mobile clients 706a-706i). For purposes of clarity, the embodiments of the present invention are described partly in conjunction with Figures 7A, 7B, and 7C. It should be noted, however, that streaming media system network 701 of Figures 7A-7C contains substantially the same components as system 300 of Figure 3. New Figures 7A-7C and system 701 are presented here, instead of again referring to Figure 3, so as to avoid unnecessarily crowding Figure 3 with the additional illustrations included in Figures 7A-7C.

Furthermore, the following discussion will present three separate cases in which the present invention is employed. First, Figure 7A, will discuss an embodiment in which two servers and a single base station are employed. Figure 7B will then be used to illustrate an embodiment in which two servers and two base stations are employed. Lastly, Figure 7C will be used to illustrate still another embodiment of the present invention is which one server and two base stations are employed. It should be understood, that in a fixed client embodiment of, for example, the depicted in Figure 7A, the present invention may assign multiple servers to transmit multiple MD bitstreams to the fixed client rather than to one or

more base stations. It should further be noted that the paths referred to in the present embodiments could be completely wired, or partially wireless. Also, the wireless part of different paths could go through different base stations (non-overlapping) or the same base station (overlapping). Even
 5 the wired parts of different path could have some overlap.

With reference now to Figure 7A, in one embodiment, each of servers 702a-702d has the computational power to deliver a certain number of MD streams to one or more base stations (e.g. base stations
 10 704a-704e) to which it has network connection of adequate network bandwidth. Base stations 704a-704e serve their respective cells. In the case of the arrangement depicted in Figure 7A, base station 704a serves mobile clients 706a and 706b. Similarly, base station 704b serves mobile clients 706c, 706d, and 706e, base station 704c serves mobile clients 706f
 15 and 706g, and base station 704e serves mobile clients 706h and 706i. Although such an arrangement is depicted in system 701 of Figure 7A, it will be understood that the various mobile clients may be served by other base stations as conditions change (e.g. a mobile client changes location, network congestion conditions change, and the like).

Referring still to Figure 7A, if there is a fairly constant demand for MD streams from mobile clients within a cell, the peak demand will differ from the average demand by only a few percentage points. In that case, a static allocation of a server's computation power can be made so that it
 20 can handle the peak demand. When the demand is only average, the server's utilization will still be very high. However because of mobility, the base station (or base stations) serving a cell may find a large population of mobile clients temporarily located within the cell. So the number of MD streams that need to be served to that cell will reach a peak
 25 during such time periods. The peak demand will differ significantly from the average demand. Since this is a transient phenomenon, all cells adjacent to the one undergoing such overload are typically not overloaded themselves. As will be discussed below, in the present embodiment, better utilization of the server's computational power will be obtained by
 30 dynamically assigning the request for MD streams to servers which are currently in the best position to serve the MD streams to the base stations and ultimately to the mobile client.
 35

In one embodiment of the present invention, a network

management protocol such as, for example, the Simple Network Management Protocol (SNMP) is used to exchange information about load in different parts of the network and on different servers. Although the following discussion will pertain to a specific network management protocol, SNMP, such reference is intended to be exemplary and is not intended to limit the inventive concepts of the present invention. That is, the various embodiments of the present invention are well suited to the use of various network management protocols other than SNMP.

With reference again to Figure 7A, the distribution of SNMP Managers (SM) and Agents (SA) among servers 702a-702d in one embodiment of the present invention is shown. SA refers to a server that runs an SNMP agent: software capable of answering valid queries in SNMP on different characteristics of the system, which in this case is the server. SM refers to a server which runs an SNMP Manager: software like network management systems that is capable of querying the SNMP agents and collecting information on network and server load. In various embodiments of the present invention, other points in the network such as, for example, routers and base stations 704a-704e, also run SNMP agents. The information from routers is also sent to neighboring SMs.

In the present embodiment, the present invention considers several factors such as the utilization of computational power on servers 702a-702d, the extent to which the streams have been cached on servers 702a-702d, and also the network bandwidth available from servers 702a-702d to the various base stations 704a-704e. Furthermore, in one embodiment of the present invention, it is assumed that placement of servers 702a-702d has been done a priori based on observed mobility patterns and demand for MD streams in different cells. It is also assumed that the designation and location of SMs and SAs has been performed a priori based on various available protocols and methods. It is further assumed that caching is being done based on similar considerations.

With reference now to Figure 8, a flow chart of steps performed in accordance with one embodiment of the present invention is shown. The methods of the present embodiment will be described in conjunction with Figure 7A and flow chart 800 of Figure 8. Although specific steps are disclosed in flow chart 800 of Figure 8, such steps are exemplary. That is, the present invention is well suited to performing various other steps or

variations of the steps recited in Figure 8. Furthermore, as mentioned above in conjunction with the description of Figure 1, portions of the present method are comprised of computer-readable and computer-executable instructions which reside, for example, in computer-usable media of a computer system. The methods of the below listed embodiments are, in some instances, comprised of computer-readable and computer-executable instructions which reside, for example, in one or more of the SMs, the SAs, the base stations, or various combinations thereof. As will be described below in detail, in the present embodiment, MD bitstreams will be provided to mobile client 706a using two servers SM 702a and SA 702b and a single base station 704a as denoted by dotted line paths 703 and 705.

At step 802, in one embodiment, a base station receives a request from a mobile client for media to be streamed to the mobile client. For purposes of illustration, the following discussion will utilize an example in which base station 704a receives a request from mobile client 706a to have media streamed thereto. The present example will further assume that the media to be streamed has been or will be encoded into two separate complimentary MD bitstreams in a manner as has been described above in detail. Furthermore, the present example also assumes that the first MD bitstream is stored at SM 702a and the second MD bitstream is stored at SA 702b. Once again, the present example is presented only for purposes of illustration of the present embodiment. That is, the present invention is well suited to the case in which: various other mobile clients request media to be streamed thereto; various other base stations receive the request for the media; and/or the requested media is encoded into other than two separate complimentary MD bitstreams stored at other than the above cited servers.

At step 804, the request for a MD stream goes to the nearest SM. In the present example, the request from mobile client 706a would be sent to SM 702a.

At step 806, the present method identifies the servers, both SMs and SAs, with a route to the base station which received the request (i.e. base station 704a in the present example). Those identified servers are considered as possible candidates to serve the MD bitstream to the mobile client. For purposes of the present example, it will be assumed that SM

702a and SA 702b, are identified as the servers with a route to base station 704a.

At steps 808, the present method then intelligently evaluates the identified servers for suitability. In one embodiment of the present invention, the present method evaluates information collected via SNMP on those servers identified as possible candidates. More specifically, in one embodiment SM 702a evaluates factors such as: the computation load of identified servers 702a and 702b; network bandwidth to base station 704a for each of identified servers 702a and 702b; and the possibility of finding the requested stream in the cache for each of identified servers 702a and 702b. Importantly, although one embodiment considers computation load, network bandwidth and potential of being cached, there are various alternatives that will be possible. One of them is the application of rules obtained by datamining of access logs (i.e. rules that map factors like classification of the content requested and the daily and seasonal variations in access characteristics to the right set of servers for serving the multiple description streams). Although such evaluation steps are recited in the present embodiment, the present invention is well suited to including various other evaluation steps and/or altering the evaluation steps mentioned above.

Importantly, it should be noted that in some embodiments of the present invention, when performing the analysis to identify the appropriate servers to use (e.g. step 808), the analysis will take into account whether two separate servers as well as two separate base stations can be used. That is, in addition to conventional metrics such as computation and bandwidth loads, etc., on the servers, an additional metric employed by some embodiments of the present invention is the diversity that can be achieved. Specifically, in some embodiment, the present invention further has a goal of maximizing the diversity (e.g. desire to have two servers sending two complementary MD bitstreams over two paths to two base stations (over two wireless links) to the client). It should be pointed out, that in one embodiment of the present invention, while intelligently evaluating the servers for assigning an MD bitstream, the diversity of paths and servers selected will also be considered to increase fault tolerance.

It should be noted, that in one embodiment of the present invention,

the intelligent evaluation process of step 808 is at least in part influenced by data sent by SNMP agents. In such an embodiment, the data is sent periodically or in response to a query from a SM. As mentioned above, although the present embodiment specifically mentions SNMP, such reference is intended to be exemplary and is not intended to limit the inventive concepts of the present invention. That is, the various embodiments of the present invention are well suited to the use of various network management protocols other than SNMP.

At step 814, based upon the above-described evaluation criterion, a server is identified as the best candidate for serving each of the MD streams. In the present example, SM 702a is identified as the best candidate for serving the first of the two MD bitstreams, and SA 702b is identified as the best candidate for serving the second of the two MD bitstreams. In summary, step 806, 807, 808 and 814 of the present embodiment comprise an analyzing process for determining the best candidate from a plurality of servers to provide respective MD bitstreams to a base station.

Referring still to step 814, a distinct and significant advantage is provided here by the present invention as compared to conventional approaches. Specifically, the present invention is able to identify and subsequently employ numerous separately located servers to provide respective complementary multiple description streams of the requested media data to the mobile client. As a result, the present embodiment provides a reliable approach for delivering the requested media data to the mobile client. Furthermore, unlike conventional approaches which supply requested data from a single source, or which repeatedly duplicate the same information onto numerous sources, the present invention is able to use multiple servers to provide requested media data to the mobile client. Hence, the present invention achieves the aforementioned reliable approach with improved efficiency as compared to conventional approaches.

With reference still to step 814, still another distinct and significant advantage is provided here by the present invention as compared to conventional approaches. Specifically, the present invention provides an intrinsic path diversity advantage without requiring inefficient complete duplication of media data to be streamed to the mobile client. That is,

using the above example of two MD streams, mobile client 706a has the potential of requesting and receiving both streams via two separate transmission paths. Thus, a single point of failure (e.g. a single server failure) does not prevent receipt by mobile client 706a of the streamed data.

5

Referring again to step 814, as will be discussed below, the present method is also well suited to an embodiment in which more than one of the plurality of MD bitstreams are provided from a single SM or SA. In one such approach, multiple base stations would be requesting the MD bitstreams, and the single server will provide the first of the MD bitstreams to the first requesting base station along a first path, and the second of the MD bitstreams to the second requesting base station along a second path.

10

15

At step 816, the present method then forwards the request to the identified best candidate. Hence, in the present example, the present method would request the first of the two MD bitstreams from SM 702a. In the present example, SA 702b would receive a request for the second of the two MD bitstreams.

20

25

At step 818, the present method causes each of the servers identified in step 814 to run an admission control process. In the present example, each of servers SM 702a and SA 702b would perform the admission control process. In the present embodiment, the outcome of the process could be any one of the three results indicated in Figure 8. That is, each of servers SM 702a and SA 702b will either: refuse admission to the requested MD bitstream (step 820); grant admission to the requested MD bitstream (step 822); or grant admission to the requested MD bitstream, but identify another stream for possible redistribution (step 824). Steps 820, 822, and 824 are each discussed separately below.

30

35

At step 820, if the server refuses admission, the present method returns to step 814. Such a condition results in another server from the set of candidates being requested to provide admission to the MD bitstream.

At step 822, if the server grants admission to the MD bitstream, that server ultimately provides the MD bitstream for the requested media data to the mobile client. Hence, in the present example, SM 702a would

provide the first of the two MD bitstreams to base station 704a, and SA 702b would provide the second of the two MD bitstreams to base station 704a.

At step 824, if the server grants admission to the MD bitstream and identifies one of its existing MD bitstreams for possible redistribution to another server, the present method proceeds to step 826. Such an outcome will occur, for example, if granting admission to the MD bitstream is possible based on average requirements of the MD bitstream, but there is significant possibility of failing to meet quality of service requirements for some MD bitstreams due to fluctuations in the bit rate. The outcome recited at step 824 may also be necessary simply because mobility patterns require occasional reassignment of MD bitstreams, and identifying these MD bitstreams when the admission control process is run is possible in the present embodiment.

Upon the completion of step 824, the present invention proceeds to step 826. At step 826, the present method then treats the server's request to reassign an existing MD bitstream to another server, in a manner similar to the approach employed when a new request from a base station is received. That is, the present embodiment proceeds to step 804 and continues from there in a manner as was described above in detail.

With reference again to Figure 8, a flow chart of steps performed in accordance with one embodiment of the present invention is shown. The methods of the present embodiment will be described in conjunction with Figure 7B and flow chart 800 of Figure 8. As will be described below in detail, in the present embodiment, MD bitstreams will be provided to mobile client 706a using two servers SM 702a and SA 702b and two base stations 704a and 704b as denoted by dotted line paths 707 and 709. In such an embodiment, two complementary MD bitstreams are sent from two separate servers and travel through two separate paths (specifically separate wired and separate wireless paths) and therefore there is no single point of failure. This contrasts with the case illustrated in Figure 7A, where there were also two separate servers and two separate wired paths, but only a single wireless path.

At step 802, in one embodiment, a base station receives a request from a mobile client for media to be streamed to the mobile client. For purposes of illustration, the following discussion will utilize an example

in which base station 704a receives a request from mobile client 706a to have media streamed thereto. The present example will further assume that the media to be streamed has been or will be encoded into two separate complimentary MD bitstreams in a manner as has been

5 described above in detail. Furthermore, the present example also assumes that the first MD bitstream is stored at SM 702a and the second MD bitstream is stored at SA 702b.

At step 804, the request for a MD stream goes to the nearest SM. In

10 the present example, the request from mobile client 706a would be sent to SM 702a.

At step 806, the present method identifies the servers, both SMs and SAs, with a route to the base station which received the request (i.e. base station 704a in the present example). Those identified servers are

15 considered as possible candidates to serve the MD bitstream to the mobile client. For purposes of the present example, it will be assumed that SM 702a and SA 702b, are identified as the servers with a route to base station 704a. In the present embodiment, however, two base stations 704a and

20 704b are identified as having capability to transmit to mobile client 706a.

At steps 808, the present method then intelligently evaluates the identified servers for suitability. In one embodiment of the present invention, the present method evaluates information collected via SNMP

25 on those servers identified as possible candidates. More specifically, in one embodiment SM 702a evaluates factors such as: the computation load of identified servers 702a and 702b; network bandwidth to base stations 704a and 704b for each of identified servers 702a and 702b, respectively; and the possibility of finding the requested stream in the cache for each of

30 identified servers 702a and 702b. Importantly, although one embodiment considers computation load, network bandwidth and potential of being cached, there are various alternatives that will be possible. One of them is the application of rules obtained by datamining of access logs (i.e. rules that map factors like classification of the content requested and the daily

35 and seasonal variations in access characteristics to the right set of servers for serving the multiple description streams). Although such evaluation steps are recited in the present embodiment, the present invention is well suited to including various other evaluation steps and/or altering the evaluation steps mentioned above.

Importantly, it should be noted that in some embodiments of the present invention, when performing the analysis to identify the appropriate servers to use (e.g. step 808), the analysis will take into account whether two separate servers as well as two separate base stations can be used. That is, in addition to conventional metrics such as computation and bandwidth loads, etc., on the servers, an additional metric employed by some embodiments of the present invention is the diversity that can be achieved. Specifically, in some embodiment, the present invention further has a goal of maximizing the diversity (e.g. desire to have two servers sending two complementary MD bitstreams over two paths to two base stations (over two wireless links) to the client). Thus, in the present embodiment to maximize diversity (along both wired and wireless links) to mobile client 706a, two servers SA 702b and SM 702a and two base stations 704a and 704b are used to send the two MD bitstreams to mobile client 706a along to completely different paths 707 and 709.

It should be noted, that in one embodiment of the present invention, the intelligent evaluation process of step 808 is at least in part influenced by data sent by SNMP agents. In such an embodiment, the data is sent periodically or in response to a query from a SM. As mentioned above, although the present embodiment specifically mentions SNMP, such reference is intended to be exemplary and is not intended to limit the inventive concepts of the present invention. That is, the various embodiments of the present invention are well suited to the use of various network management protocols other than SNMP.

At step 814, based upon the above-described evaluation criterion, a server is identified as the best candidate for serving each of the MD streams. In the present example, SM 702a is identified as the best candidate for serving the first of the two MD bitstreams to base station 704a, and SA 702b is identified as the best candidate for serving the second of the two MD bitstreams to base station 704b. In summary, step 806, 807, 808 and 814 of the present embodiment comprise an analyzing process for determining the best candidates from a plurality of servers to provide respective MD bitstreams to a plurality of base stations.

Referring still to step 814, a distinct and significant advantage is

provided here by the present invention as compared to conventional approaches. Specifically, the present invention is able to identify and subsequently employ numerous separately located servers and base stations to provide respective complementary multiple description streams of the requested media data to the mobile client. As a result, the present embodiment provides a reliable approach for delivering the requested media data to the mobile client. Furthermore, unlike conventional approaches which supply requested data from a single source, or which repeatedly duplicate the same information onto numerous sources, the present invention is able to use multiple servers and base stations to provide requested media data to the mobile client. Hence, the present invention achieves the aforementioned reliable approach with improved efficiency as compared to conventional approaches.

With reference still to step 814, still another distinct and significant advantage is provided here by the present invention as compared to conventional approaches. Specifically, the present invention provides an intrinsic path diversity advantage without requiring inefficient complete duplication of media data to be streamed to the mobile client. That is, using the above example of two MD streams, mobile client 706a has the potential of requesting and receiving both streams via two completely separate transmission paths 707 and 709. Thus, a single point of failure (e.g. a single server failure or single base station failure) does not prevent receipt by mobile client 706a of the streamed data.

Referring again to step 814, as will be discussed below, the present method is also well suited to an embodiment in which more than one of the plurality of MD bitstreams are provided from a single SM or SA. In one such approach, multiple base stations would be requesting the MD bitstreams, and the single server will provide the first of the MD bitstreams to the first requesting base station along a first path, and the second of the MD bitstreams to the second requesting base station along a second path.

At step 816, the present method then forwards the request to the identified best candidate. Hence, in the present example, the present method would request the first of the two MD bitstreams from SM 702a. In the present example, SA 702b would receive a request for the second of the

two MD bitstreams.

At step 818, the present method causes each of the servers identified in step 814 to run an admission control process. In the present example, each of servers SM 702a and SA 702b would perform the admission control process. In the present embodiment, the outcome of the process could be any one of the three results indicated in Figure 8. That is, each of servers SM 702a and SA 702b will either: refuse admission to the requested MD bitstream (step 820); grant admission to the requested MD bitstream (step 822); or grant admission to the requested MD bitstream, but identify another stream for possible redistribution (step 824). Steps 820, 822, and 824 are each discussed separately below.

At step 820, if the server refuses admission, the present method returns to step 814. Such a condition results in another server from the set of candidates being requested to provide admission to the MD bitstream.

At step 822, if the server grants admission to the MD bitstream, that server ultimately provides the MD bitstream for the requested media data to the mobile client. Hence, in the present example, SM 702a would provide the first of the two MD bitstreams to base station 704b, and SA 702b would provide the second of the two MD bitstreams to base station 704a.

At step 824, if the server grants admission to the MD bitstream and identifies one of its existing MD bitstreams for possible redistribution to another server, the present method proceeds to step 826. Such an outcome will occur, for example, if granting admission to the MD bitstream is possible based on average requirements of the MD bitstream, but there is significant possibility of failing to meet quality of service requirements for some MD bitstreams due to fluctuations in the bit rate. The outcome recited at step 824 may also be necessary simply because mobility patterns require occasional reassignment of MD bitstreams, and identifying these MD bitstreams when the admission control process is run is possible in the present embodiment.

Upon the completion of step 824, the present invention proceeds to step 826. At step 826, the present method then treats the server's request to reassign an existing MD bitstream to another server, in a manner

similar to the approach employed when a new request from a base station is received. That is, the present embodiment proceeds to step 804 and continues from there in a manner as was described above in detail.

With reference again to Figure 8, a flow chart of steps performed in accordance with one embodiment of the present invention is shown. The methods of the present embodiment will be described in conjunction with Figure 7C and flow chart 800 of Figure 8. As will be described below in detail, in the present embodiment, MD bitstreams will be provided to mobile client 706a using one servers SM 702a and two base stations 704a and 704b as denoted by dotted line paths 711 and 713. In such an embodiment, two complementary MD bitstreams are sent from a single server and travel through two separate paths (specifically separate wired and separate wireless paths). This contrasts with the case illustrated in Figure 7A, where there were also two separate servers and two separate wired paths, but only a single wireless path, and with the case illustrated in Figure 7B where there were also two separate servers and two separate base stations.

At step 802, in one embodiment, a base station receives a request from a mobile client for media to be streamed to the mobile client. For purposes of illustration, the following discussion will utilize an example in which base station 704a receives a request from mobile client 706a to have media streamed thereto. The present example will further assume that the media to be streamed has been or will be encoded into two separate complimentary MD bitstreams in a manner as has been described above in detail. Furthermore, the present example also assumes that the first MD bitstream and the second MD bitstream is stored at SM 702a.

At step 804, the request for a MD stream goes to the nearest SM. In the present example, the request from mobile client 706a would be sent to SM 702a.

At step 806, the present method identifies the servers, both SMs and SAs, with a route to the base station which received the request (i.e. base station 704a in the present example). Those identified servers are considered as possible candidates to serve the MD bitstream to the mobile client. For purposes of the present example, it will be assumed that SM

702a and SA 702b, are identified as the servers with a route to base station 704a but that only SM 702a is to be used. In the present embodiment, however, two base stations 704a and 704b are identified as having capability to transmit to mobile client 706a.

5

At steps 808, the present method then intelligently evaluates single server 702a for suitability. In one embodiment of the present invention, the present method evaluates information collected via SNMP on those servers identified as possible candidates. More specifically, in one
 10 embodiment SM 702a evaluates factors such as: the computation load of identified server 702a; network bandwidth to base stations 704a and 704b for identified server 702a, respectively; and the possibility of finding the requested stream in the cache for identified server 702a. Importantly, although one embodiment considers computation load, network
 15 bandwidth and potential of being cached, there are various alternatives that will be possible. One of them is the application of rules obtained by datamining of access logs (i.e. rules that map factors like classification of the content requested and the daily and seasonal variations in access characteristics to the right server for serving the multiple description
 20 streams). Although such evaluation steps are recited in the present embodiment, the present invention is well suited to including various other evaluation steps and/or altering the evaluation steps mentioned above.

25

Importantly, it should be noted that in some embodiments of the present invention, when performing the analysis to identify the appropriate server to use (e.g. step 808), the analysis will take into account whether two separate base stations can be used. That is, in addition to conventional metrics such as computation and bandwidth loads, etc., on
 30 the server, an additional metric employed by some embodiments of the present invention is the diversity that can be achieved. Specifically, in some embodiment, the present invention further has a goal of maximizing the diversity (e.g. desire to have the single server send two complementary MD bitstreams over two paths to two base stations (over
 35 two wireless links) to the client). Thus, in the present embodiment to maximize diversity (along both wired and wireless links) to mobile client 706a, two base stations 704a and 704b are used to send respective ones of the MD bitstreams to mobile client 706a along to completely different paths
 711 and 713.

It should be noted, that in one embodiment of the present invention, the intelligent evaluation process of step 808 is at least in part influenced by data sent by SNMP agents. In such an embodiment, the data is sent
5 periodically or in response to a query from a SM. As mentioned above, although the present embodiment specifically mentions SNMP, such reference is intended to be exemplary and is not intended to limit the inventive concepts of the present invention. That is, the various embodiments of the present invention are well suited to the use of various
10 network management protocols other than SNMP.

At step 814, based upon the above-described evaluation criterion, a server, SM 702a, is identified as the best candidate for serving the first of the two MD bitstreams to base station 704a and the second of the two MD
15 bitstreams to base station 704b. In summary, step 806, 807, 808 and 814 of the present embodiment comprise an analyzing process for determining the best candidate from a plurality of servers to provide MD bitstreams to a plurality of base stations.

Referring still to step 814, a distinct and significant advantage is provided here by the present invention as compared to conventional approaches. Specifically, the present invention is able to identify and subsequently employ numerous separately located base stations to provide
20 respective complementary multiple description streams of the requested media data to the mobile client. As a result, the present embodiment provides a reliable approach for delivering the requested media data to the mobile client. Furthermore, unlike conventional approaches which supply requested data from a single source via a single path, or which repeatedly duplicate the same information onto numerous sources, the
25 present invention is able to use multiple base stations to provide requested media data to the mobile client. Hence, the present invention achieves the aforementioned reliable approach with improved efficiency as compared to conventional approaches.

With reference still to step 814, still another distinct and significant advantage is provided here by the present invention as compared to conventional approaches. Specifically, the present invention provides an intrinsic path diversity advantage without requiring inefficient complete duplication of media data to be streamed to the mobile client. That is,
35

using the above example of two MD streams, mobile client 706a has the potential of requesting and receiving both streams via two completely separate transmission paths 711 and 713. Thus, a single point of failure (e.g. a single base station failure) does not prevent receipt by mobile client
5 706a of the streamed data.

At step 816, the present method then forwards the request to the identified best candidate. Hence, in the present example, the present method would request both of the two MD bitstreams from SM 702a.

10

At step 818, the present method causes the server identified in step 814 (SM 702a) to run an admission control process. In the present embodiment, the outcome of the process could be any one of the three results indicated in Figure 8. That is, server SM 702a will either: refuse
15 admission to the requested MD bitstream (step 820); grant admission to the requested MD bitstream (step 822); or grant admission to the requested MD bitstream, but identify another stream for possible redistribution (step 824). Steps 820, 822, and 824 are each discussed separately below.

20

At step 820, if the server refuses admission, the present method returns to step 814. Such a condition results in another server from the set of candidates being requested to provide admission to the MD bitstream.

25

At step 822, if the server grants admission to the MD bitstream, that server ultimately provides the MD bitstream for the requested media data to the mobile client. Hence, in the present example, SM 702a would provide the first of the two MD bitstreams to base station 704a and the second of the two MD bitstreams to base station 704b.

30

At step 824, if the server grants admission to the MD bitstream and identifies one of its existing MD bitstreams for possible redistribution to another server, the present method proceeds to step 826. Such an outcome will occur, for example, if granting admission to the MD bitstream is
35 possible based on average requirements of the MD bitstream, but there is significant possibility of failing to meet quality of service requirements for some MD bitstreams due to fluctuations in the bit rate. The outcome recited at step 824 may also be necessary simply because mobility patterns require occasional reassignment of MD bitstreams, and identifying these

MD bitstreams when the admission control process is run is possible in the present embodiment.

Upon the completion of step 824, the present invention proceeds to step 826. At step 826, the present method then treats the server's request to reassign an existing MD bitstream to another server, in a manner similar to the approach employed when a new request from a base station is received. That is, the present embodiment proceeds to step 804 and continues from there in a manner as was described above in detail.

Thus, the present invention provides a method and system for streaming media to fixed clients and/or mobile clients. The present invention further provides a method and system for streaming media to fixed clients and/or mobile clients wherein the method and system provides increased reliability and efficiency over conventional systems.

The foregoing descriptions of specific embodiments of the present invention have been presented for purposes of illustration and description. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed, and obviously many modifications and variations are possible in light of the above teaching. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the Claims appended hereto and their equivalents.